Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 3439 – 3444

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineeringand Information Science

A Novel Automatic Image Annotation Method Based on Multi-instance Learning

Shunle Zhu[a] ,Xiaoqiu Tan[a],a*

School of mathematics, physics and information, Zhejiang Ocean University, Zhoushan, 316000, China

Abstract

Automatic image annotation (AIA) is the bridge of high-level semantic information and the low-level feature. AIA is an effective method to resolve the problem of "Semantic Gap". According to the intrinsic character of AIA, which is many regions contained in the annotated image, AIA Based on the framework of multi-instance learning (MIL) is proposed in this paper. Each keyword is analyzed hierarchically in low-granularity-level under the framework of MIL. Through the representative instances are mined, the semantic similarity of images can be effectively expressed and the better annotation results are able to be acquired, which testifies the effectiveness of the proposed annotation method.

Keywords: automatic image annotation; multi-instance learning; representative instances; semantic similarity;

1. Introduction

With the development of multimedia and network technology, image data has been becoming more common rapidly. Facing a mass of image resource, content based image retrieval (CBIR), a technology to organize, manage and analyze these resource efficiently, is becoming a hot point. However, under the limitation of "semantic gap", that is, the underlying vision features, such as color, texture, and shape, can not reflect and match the query attention completely, CBIR confronts the unprecedented challenge.

In recent years, newly proposed automatic image annotation (AIA) keeps focus on erecting a bridge between high-level semantic and low-level features, which is an effective approach to solve the above

* Shunle Zhu. Tel.: 13567658387; fax: 0580-2627253.
E-mail address: zhehaiyang@126.com

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2011.08.644

mentioned semantic gap. Since 1999 co-occurrence model proposed by Morris etc., the research of automatic image annotation was initiated[1]. In [2], translation model was developed to annotate image automatically based on an assumption that keywords and vision features were different language to describe the same image. Similar to [2], literature [3] proposed Cross Media Relevance Model (CMRM) where the vision information of each image was denoted as blob set which is to manifest the semantic information of image. However, blob set in CMRM was erected based on discrete region clustering which produced a loss of vision features so that the annotation results were too perfect. In order to compensate for this problem, a Continuous-space Relevance Model (CRM) was proposed in [4]. Furthermore, in [5] Multiple-Bernoulli Relevance Model was proposed to improve CMRM and CRM.

Despite variable sides in the above mentioned methods, the core idea based on automatic image annotation is identical. The core idea of automatic image annotation applies annotated images to erect a certain model to describe the potential relationship or map between as keywords and image features which is used to predict unknown annotation images. Even if previous literatures achieved some results from variable sides respectively, semantic description of each keyword has not been defined explicitly in them. For this end, on the basis of investigating the characters of the automatic image annotation, i.e. images annotated by keywords comprise multiple regions; automatic image annotation is regarded as a problem of multi instance learning. The proposed method analyzes each keyword in multi-granularity hierarchy to reflect the semantic similarity so that the method not only characterizes semantic implication accurately but also improves the performance of image annotation which verifies the effectiveness of our proposed method.

This article is organized as follows: section 1 introduces automatic image annotation briefly; automatic image annotation based on multi-instance learning framework is discussed in detail in section 2; and experimental process and results are described in section 3; section 4 summaries and discusses the future research briefly.

2. Automatic Image Annotation in the framework of Multi-instance Learning

In the previous learning framework, a sample is viewed as an instance, i.e. the relationship between samples and instances is one-to-one, while a sample may contain more instances, this is to say, the relationship between samples and instances is one-to-many. Ambiguities between training samples of multi-instance learning differ from ones of supervised learning, unsupervised learning and reinforcement learning completely so that the previous methods hardly solve the proposed problems. Owing to its characteristic features and wide prospect, multi-instance learning is absorbing more and more attentions in machine learning domain and is referred to as a newly learning framework[7]. The core idea multi-instance learning is that the training sample set consists of concept-annotated bags which contain unannotated instances. The purpose of multi-instance learning is to assign a conceptual annotation to bags beyond training set by learning from training bags. In general, a bag is annotated a Positive if and only if at least one instance is labeled Positive, otherwise the bag is annotated as Negative.

2.1. Framework of Image Annotation of Multi-instance Learning

According to the above-mentioned definition of the multi-instance learning, namely, a Positive bag contain at least a positive instance, we can draw a conclusion that positive instances should be distributed much more than negative instances in Positive bags. This conclusion shares common properties with DD algorithm [8] in multi-instance learning domain. If some point can represent the more semantic of a specified keyword than any other point in the feather space, no less than one instance in positive bags should be close to this point while all instances in negative bags will be far away from this point. In the

proposed methods, we take into consideration each semantic keyword independently. Even if a part of useful information will be lost neglecting the relationship between keywords, various keywords from each image are used to computing the similarities between images so that the proposed methods can represent the semantic similarity of image effectively in low- granularity. In the following sections, each keyword will be analyzed and applied in local level so that irrelevant information with keywords will be eliminated to improve the precision of representation of the semantic of keywords. Firstly, keywords w, including Positive and Negative bags, are collected, and the area surrounded by Positive bags are obtained by clustering adaptively. Secondly, this cluster is viewed as Positive set of w which contains most items than other clusters and is farthest from Negative bags. Thirdly, Gaussian Mixture Model (GMM) is used to learn the semantic of w . Finally, the images can be annotated automatically based on the posterior probability of each keyword of images according to the probability of image in GMM by using Bayesian estimation. Figure 1 illustrates this process.

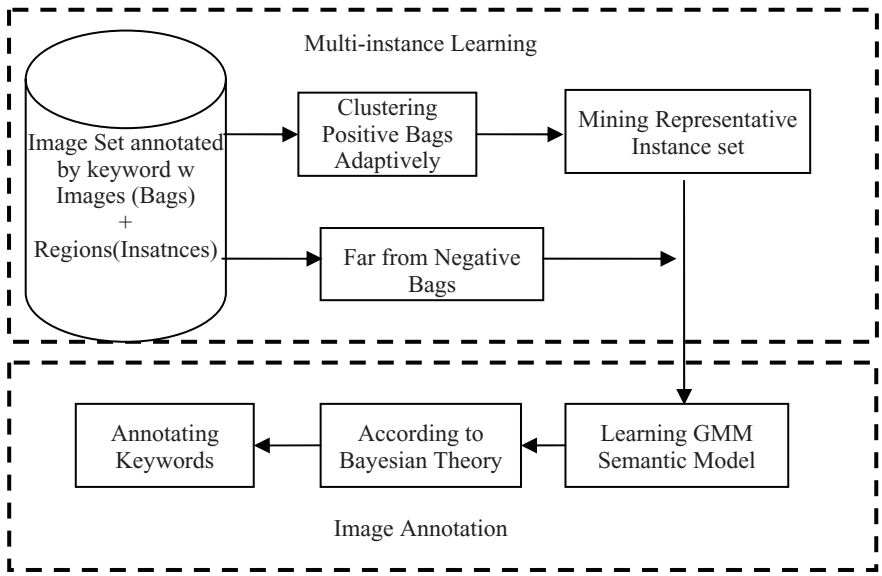


Fig.1. The framework of automatic image annotation based multi-instance learning

2.2. Automatic Image Annotation

In convenience, we firstly put forward some symbols. w is denoted as a semantic keyword, $\bar{X}=\{X_k|k=1,...,N\}$ as a set of training samples, where N is the number of training samples; $S=\{x_1^*,\mathrm{L}\ ,x_n^*\}$ as a set of representative instances after adaptively clustering, where x_n^* is the nth item in a clusters. Therefore, GMM is constructed to describe semantic concept of w, i.e. GMM is used to estimate the distribution of each keyword of feature space to erect the one-to-one map from keywords to vision feature. Note that the superiority of GMM lies in producing a smooth estimation for any density distribution which can reflect the feature distribution of semantic keywords effectively by non-parameter density estimating.

For a specified keyword w, GMM represents its vision feature distribution, $p(x|w)$ is defined as follows:

$$p(x|w)=\sum_{i=1}^{M}\pi_i N(x|\mu_i,\Sigma_i) \tag{1}$$

Where $N(x|\mu_i,\Sigma_i)$ represents the Gussian distribution of i$^{\text{th}}$ component, μ_i and Σ_i are the corresponding mean and variance respectively, π_i is weight of the i$^{\text{th}}$ component, reflecting its significance, and $\sum_{i=1}^{M}\pi_i=1$,

M is the number of components. Each component represents a cluster in feature space, reflecting a vision feature of w. In each component, the conditional probability density of low-level vision feature vector x can be computed as follows:

$$N(x;\mu_i;\Sigma_i)=\frac{1}{\sqrt{(2\pi)^d\,|\Sigma_i|}}\exp\left\{-\frac{1}{2}(x-\mu_i)^T\,\Sigma_i^{-1}(x-\mu_i)\right\} \quad (2)$$

Where d is the dimension of feature vector x. The parameters of GMM are estimated by EM method which is maximum likelihood estimation for distribution parameters from incomplete data. EM consists of two steps, expectation step, E-step, and maximum step, M-step, which are executed alternately until convergence after multiple iteration. Assuming that the keyword w can produce N_w representative instances, $\theta_i=(\mu_i,\Sigma_i)$ represents mean and co-variance of the i[th] Gussian component. Intuitively, different semantic keywords should represent different vision features and the numbers of components are not identical with each other in general so that an adaptive value of M can be obtained based on Minimum Description Length (MDL)[9].

The proposed method extracts semantic clustering sets from training images which are used to construct GMM in which each component represents some vision feature of a specified keyword. From the perspective of semantic mapping, the proposed model described the one-to-many relationship between keywords and the corresponding vision features. The extracted semantic clustering set can reflect the semantic similarity between instances and keywords. According to the above methods, a GMM is constructed for each keyword respectively to describe the semantic of the keyword. And then, for a specified image to be annotated $X=\{x_1,\ldots,x_m\}$, where x_m is denoted as the m[th] separated region, the probability of keyword w is computed according to formula (3).

$$p(w|X)\propto\prod_{i=1}^{m}p(x_i|w) \quad (3)$$

Finally, the image X is annotated according to 5 keywords of greatest posterior probabilities.

3. Experimental Results and Analysis

For comparison with other image annotation algorithms fairly, COREL[2], a widely used image data set, is selected in our experimental process. This image set consists of 5000 images, 4500 images from which are used as training samples, the rest 500 images as test samples. 1 through 5 keywords is extracted to annotate an image, so in all 371 keywords exists in dataset. In our experiments, each image is divided 10 regions using Normalized Cut segment technology [6]. 42,379 regions are produced in all for a whole image data set, and then, these regions are clustered to 500 groups each of which is called a blob. For each region, 36-demension features, such as color, shape, location etc. are considered like literature [2].

In order to measure the performances of various image annotation methods, we adopt the same evaluation metrics as literature [5], some popular indicators in automatic image annotation and image retrieval. Precision is referred as the ratio of the times of correct annotation in relation to all the times of annotation, while recall is referred as the ratio of the times of correct annotation in relation to all the positive samples. The detailed definitions are as follows:

$$precision=\frac{B}{A} \quad (4)$$

$$recall=\frac{B}{C} \quad (5)$$

Where A is the number of images annotated by some keyword; B is the number of images annotated correctly; C is the number of images annotated by some keyword in the whole data set. As a tradeoff between the above indicators, the geometric mean of them is adopted widely, namely:

$$F = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right) \quad (6)$$

Moreover, we take a statistics of the number of keywords annotated correctly which are used to annotate an image correctly at least. The statistical value reflects the coverage of keywords in our proposed methods, denoted by "NumWords".

3.1. Experimental Results

Figure 2 shows that the annotated results of the proposed method, MIL Annotation, keep rather a high consistent with the ground truth. This fact verifies the effectiveness of our proposed methods.

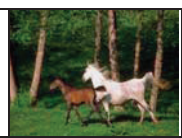
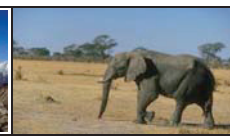
Test Images					
Ground Truth	horse grass tree animal	winter ground mountain sky	elephant ground tree animal sky	bus vehicle ground building	building tree grass sky
MIL Annotation	horse animal tree field grass	mountain winter sand ground sky	elephant animal ground tree sky	bus vehicle tree building track	building tree horse grass sky

Fig.2. Illustrations of annotation results of MIL Annotation

3.2. Annotation Results of MIL Annotation

Table 1 and Table 2 show that compare the average performance between our proposed method and some traditional annotation models such as COM[1], TM[2], CMRM[3], CRM[4] and MBRM[5], on COREL image data set. In experiments, 263 keywords are concerned.

Table 1. The performances of various annotation model on COREL

Models	COM	TM	CMRM	CRM	MBRM	MIL
Num Words	19	49	66	107	122	124
Results on 263 keywords						
Average Precision	0.03	0.06	0.10	0.16	0.24	0.20
Average Recall	0.02	0.04	0.09	0.19	0.25	0.22

Table 2. The comparison of F-measure between various models

Models	COM	TM	CMRM	CRM	MBRM	MIL
F-Measurement	0.024	0.048	0.095	0.174	0.245	0.211

From Table 1 and Table 2, we can know that the annotation performance of the proposed method outperforms other models in two keyword set, and the proposed method has a significant improvement relation to existing algorithms in average precision, average recall F-measure and "NumWords". Specifically, MIL annotation can obtain a significant improvement over COM, TM, CMRM and CRM; in existing probability-based image annotation models, MBRM can get a best annotation performance which is equivalent to the performance of MIL annotation.

4. Conclusions

Analyzing the properties of automatic image annotation deeply can know it can be viewed as a multi-instance learning problem so that we proposed a method to annotated images automatically based on multi-instance learning. Each keyword is analyzed independently to guarantee more effective semantic similarity in low-granularity. And then, under the frame of multi-instance learning, each keyword is further analyzed in various hierarchies. Irrelevant information with keywords will be eliminated to improve the precision of representation of the semantic of keywords by mapping keywords to corresponding region. Experimental results demonstrated the effectiveness of MR-MIL.

References

[1] Mori Y, Takahashi H, Oka R. Image-to-word transformation based on dividing and vector quantizing images with words. In: Proc. of Intl. Workshop on *Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, Orlando, Oct. 1999.

[2] Duygulu P, Barnard K, Freitas N, Forsyth D. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proc. of European Conf. on *Computer Vision (ECCV'02)*, Copenhagen, Denmark, May. 2002: 97-112.

[3] Jeon J, Lavrenko V, Manmatha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proc. of Int. ACM SIGIR Conf. on *Research and Development in Information Retrieval (ACM SIGIR'03)*, Toronto, Canada, Jul. 2003: 119-126.

[4] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures. In: Proc. Of *Advances in Neural Information Processing Systems (NIPS'03)*, 2003.

[5] Feng S, Manmatha R, Lavrenko V. Multiple bernoulli relevance models for image and video annotation. In: Proc. of IEEE Int. Conf. on *Computer Vision and Pattern Recognition (CVPR'04)*, Washington DC, USA, Jun. 2004: 1002-1009.

[6] Shi J, Malik J. Normalized cuts and image Segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[7] Maron O. Learning from ambiguity. *Department of Electrical Engineering and Computer Science*, MIT, PhD dissertation. 1998.

[8] Maron O, Lozano P T. A framework for multiple-instance learning. In: Proc. of *Advances in Neural Information Processing Systems (NIPS'98)*, Pittsburgh, USA, Oct. 1998: 570-576.

[9] Li J, Wang J. Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Trans. *On Pattern Analysis and Machine Intelligence*, 2003, 25(9): 1075 - 1088.